VIBE PROJECT

Virtual Biomedical and STEM/STEAM Education

2021-1-HU01-KA220-HED-000032251

Funded by the European Union

Erasmus+

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS

U.PORTO

Politechnika Śląska

DEX innovation centre

VIBE PROJECT

PATTERN RECOGNITION IN BIOMEDICAL ENGINEERING

FEATURE EXTRACTION AND SELECTION

Source: https://www.creativefabrica.com/pl/product/its-not-a-bug-its-a-feature/

Introduction to Feature Engineering

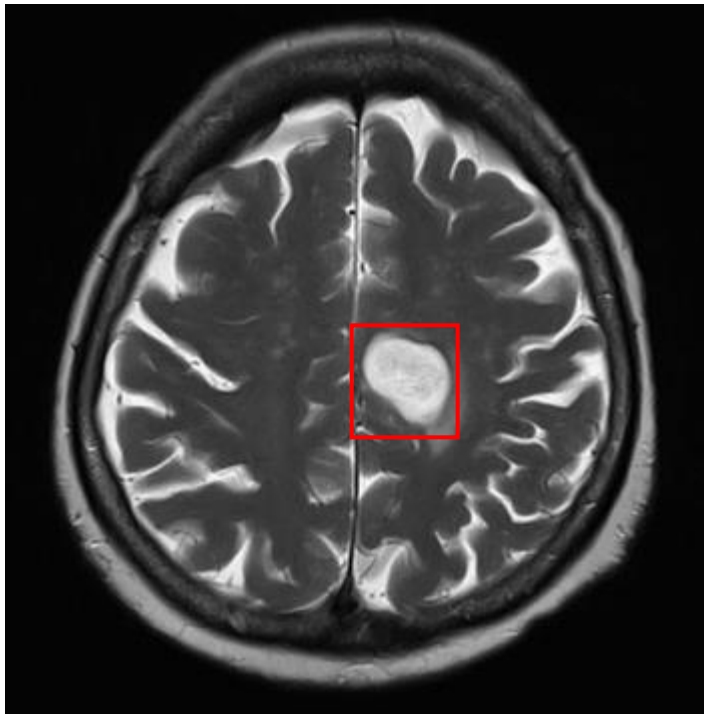
- Definition: Feature engineering involves creating and selecting the most relevant features for predictive models.
- Role in Biomedical Engineering: Critical for interpreting complex data like images, signals, and genetic information.
- Types of Biomedical Data: Imaging, Signals, Genomics, etc.

Feature Extraction

- Definition: Process of transforming raw data into informative features.
- Techniques:
  - Signal Processing for biomedical signals (e.g., wavelet transforms for ECG)
  - Texture Analysis for Images (e.g., GLCM for MRI)
  - Dimensionality Reduction (e.g., PCA for high-dimensional data)
- Examples: Specific features extracted from ECG, MRI, and genetic data.

Types of Features

- Structural Features: Shape, size, and location information.



- Number of pixels (area)
- Bounding box
- Center of mass (centroid)
- Circularity
- ...

Source: https://www.shutterstock.com/pl/image-photo/axial-cut-magnetic-resonance-image-mri-1565500459

Types of Features

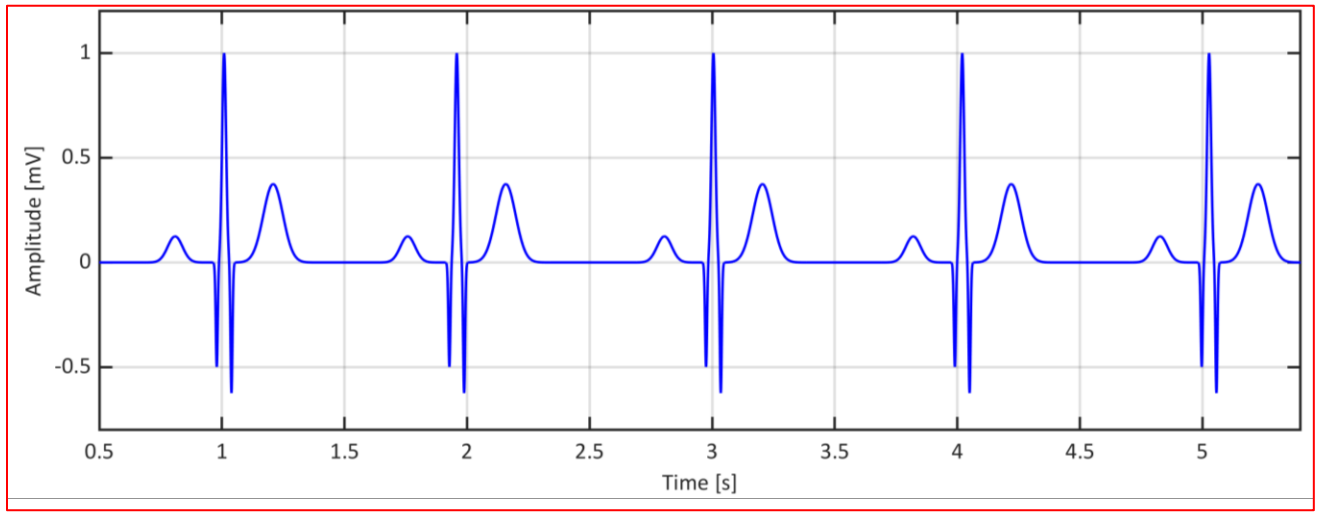
- Statistical Features: Mean, variance, skewness, etc.

N	1	2	3
White blood cells [no/µl]	5600	9000	4800
Red blood cells [no/µl]	4.8x106	4.1x106	4.6x106
Platelet count [no/µl]	142,000	148,000	151,000
Hemoglobin [g/dl]	12.2	13.4	12.5
Glucose [mg/dl]	71	74	72

- $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- …

Types of Features

- Temporal Features: Patterns over time (e.g., ECG signals).

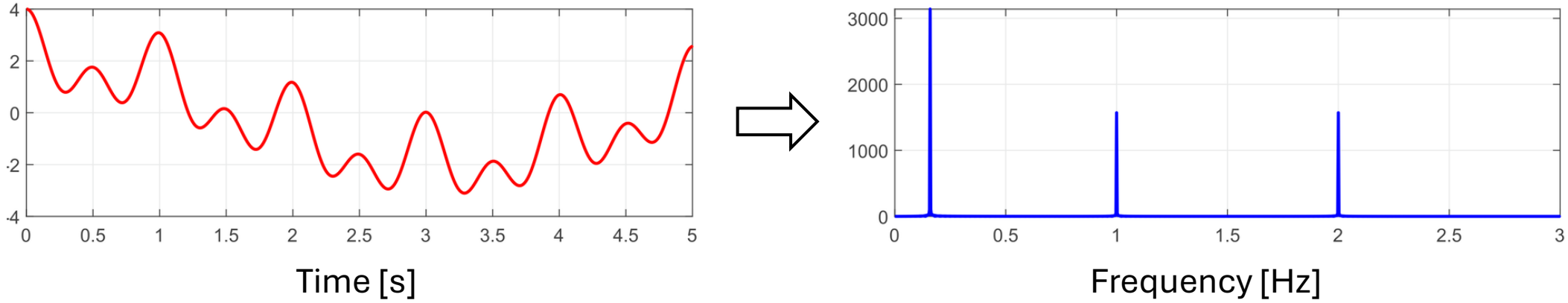


HRV = 0.05 s

Source: Original work by the author.

Types of Features

- Frequency-Domain Features: Derived from signal transformations



$$X(\omega) = \sum_{n=-\infty}^{n=+\infty} x(n)e^{-j\omega n}, \omega \in (-\pi, \pi)$$

Source: Original work by the author.

Feature importance analysis

Analysis of Variance (ANOVA)

$$F = \frac{Between - group\ variance}{Within - grop\ variance} = \frac{\sum_{k=1}^{K} n_k (\overline{x}_k - \overline{x})^2 / (K-1)}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(x_{i,k} - \overline{x}_k \right)^2 / (N-K)},$$

Where:

- K: number of groups (categories of the target variable)
- N: total number of observations
- n_k: number of observations in the k-th group
- $\overline{x}$: overall mean
- $\overline{x}_k$: mean of the k-th group

Minimum Redundancy Maximum Relevance (MRMR)

- Relevance : $I(X_i; Y) = \sum_{x_i \in X_i} \sum_{y \in Y} P(x_i, y) \log \frac{P(x_i, y)}{P(x_i)P(y)}$
- Redundancy: $I(X_i; X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$
- MRMR Criterion:

$$\text{MRMR} = \max \left(\frac{1}{|S|} \sum_{X_i \in S} I(X_i; Y) - \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i; X_j) \right)$$

Feature selection

Wrapper Methods

Forward Feature Selection

1. Rank features based on their importance
2. Train the model on small number of features
3. Iteratively add features to a model until the desired performance is reached

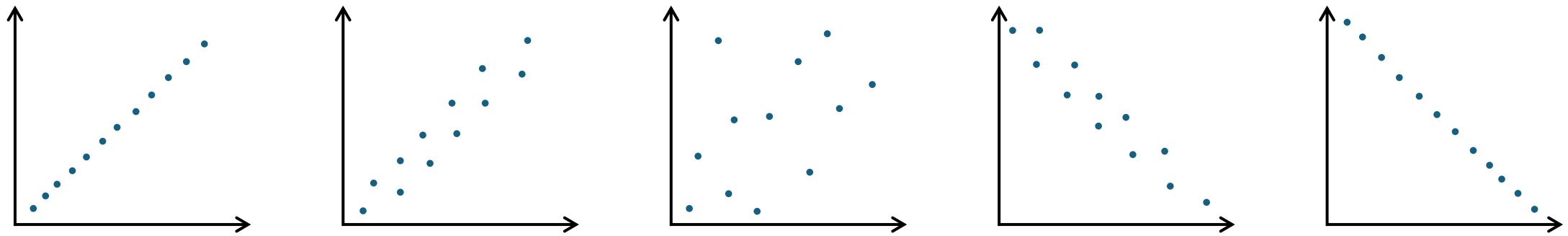
Wrapper Methods

Backward Feature Elimination

1. Train model on all features
2. Rank features based on their importance
3. Eliminate the least important features and repeat until the desired numer of features is reached

Filter Methods

Correlation analysis



r = 1
Perfect positive correlation

r = 0.8
Medium positive correlation

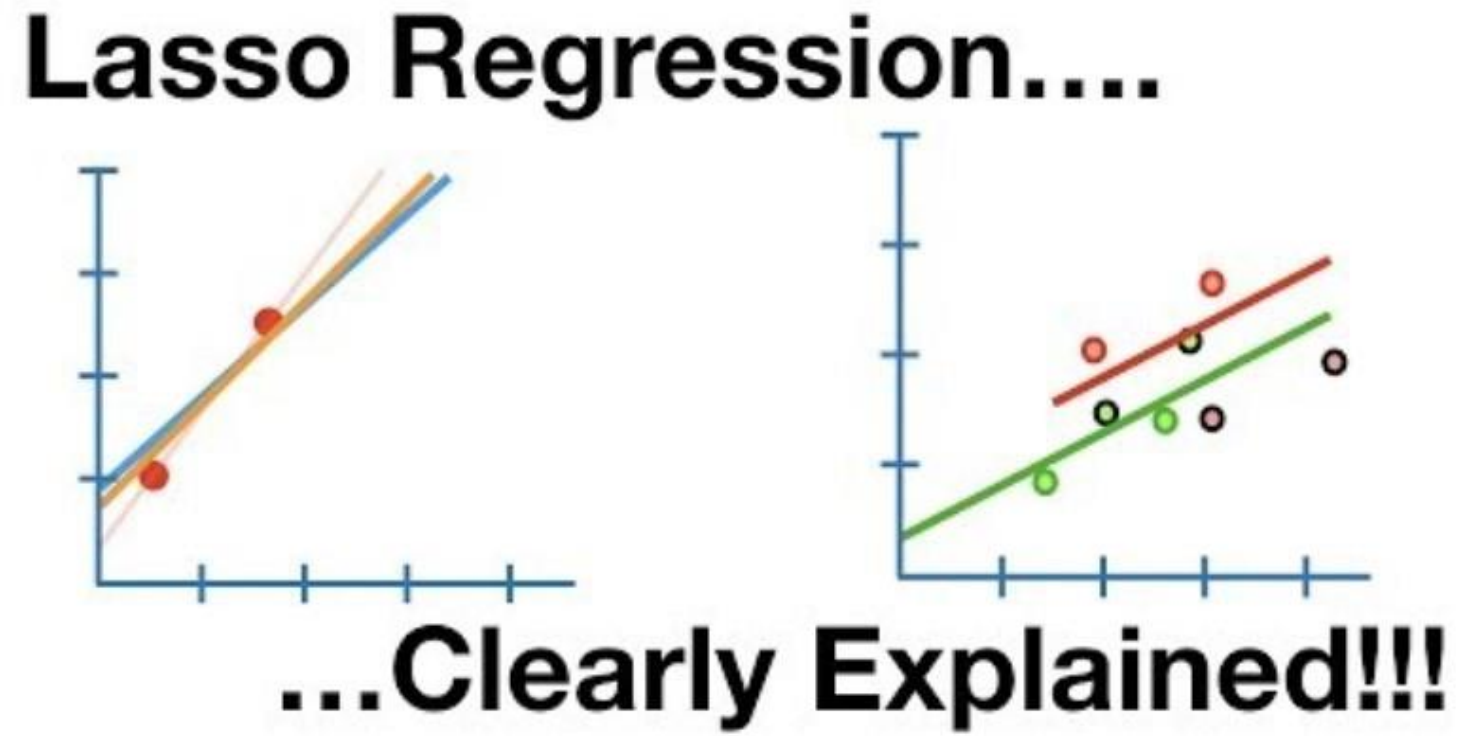
r = 0
No correlation

r = -0.8
Medium negative correlation

r = -1
Perfect negative correlation

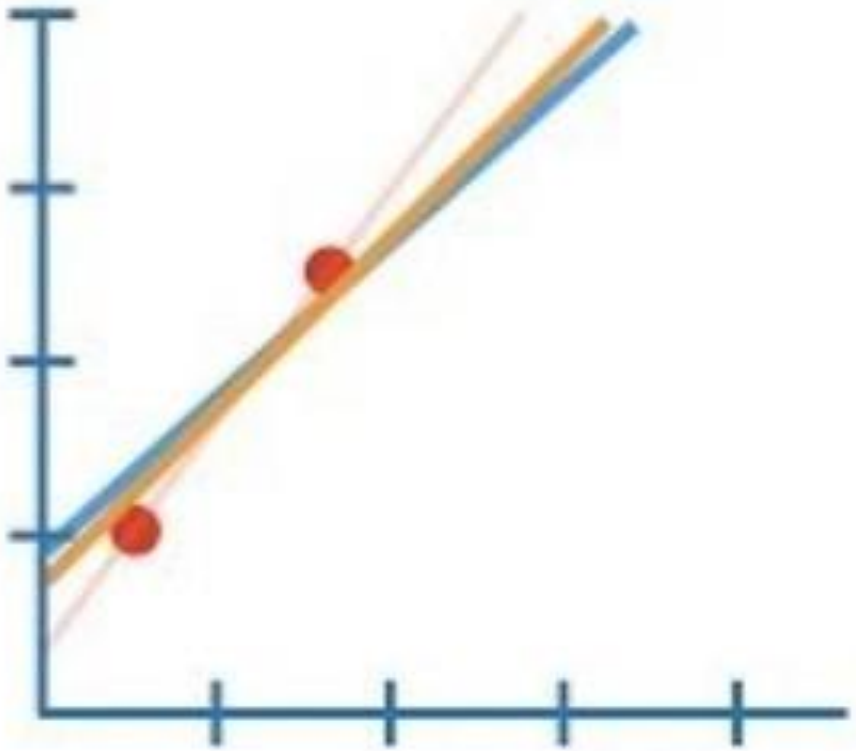
Source: Original work by the author.

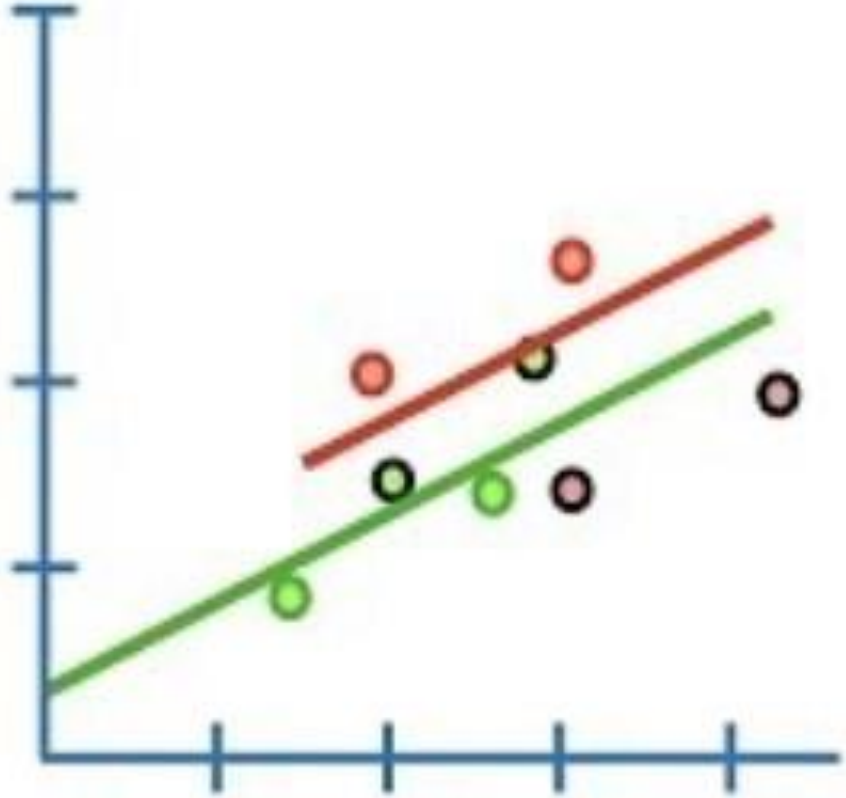
Embedded Methods



Source: https://www.youtube.com/watch?v=NGf0voTMlcs

Lasso Regression....





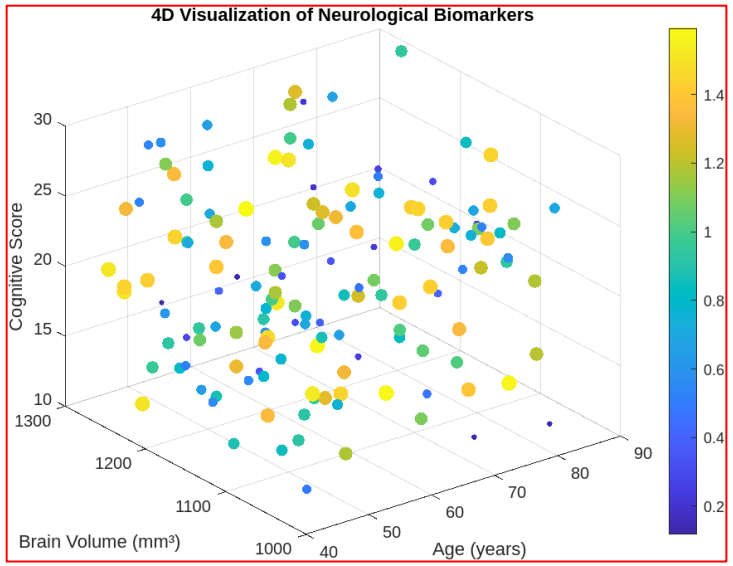
Challenges and Considerations in Biomedical Feature Engineering

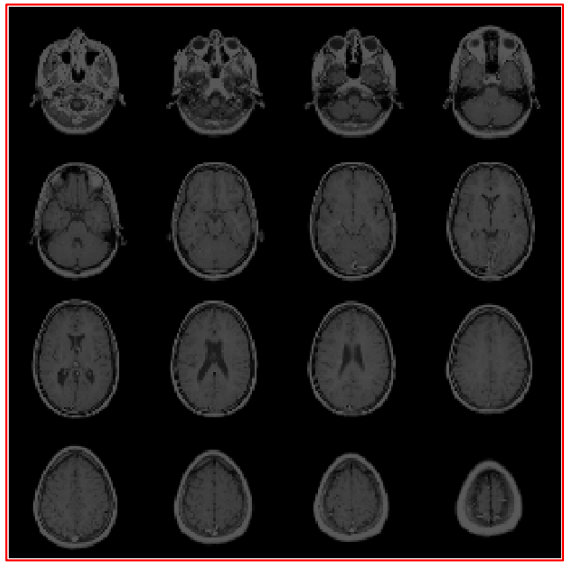
Data Variability

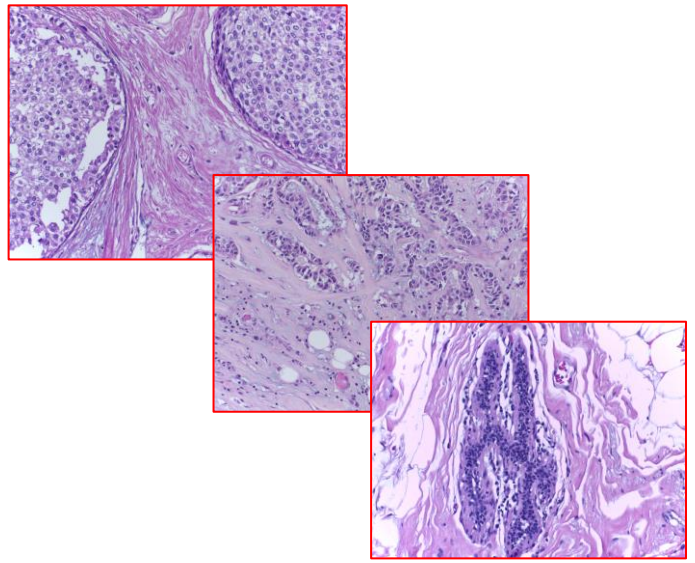
- Biological differences
- Calibration
- Equipment
- Human error
- …

High Dimensionality & Small Sample Sizes

- Common in biomedical data, leading to overfitting risks.



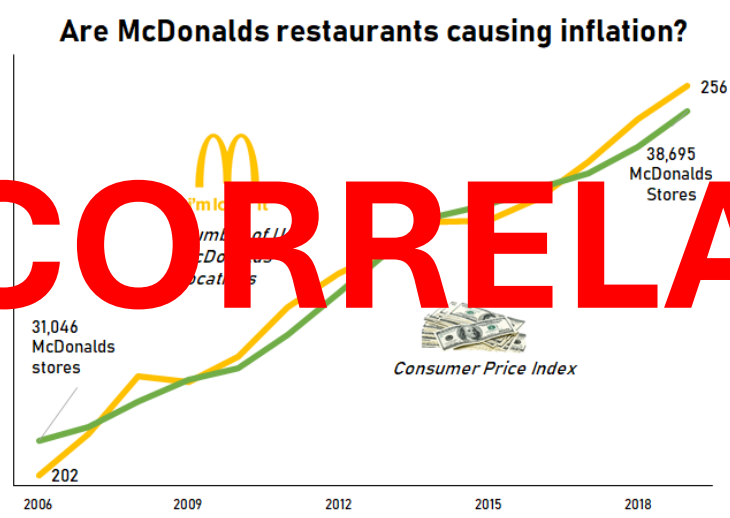


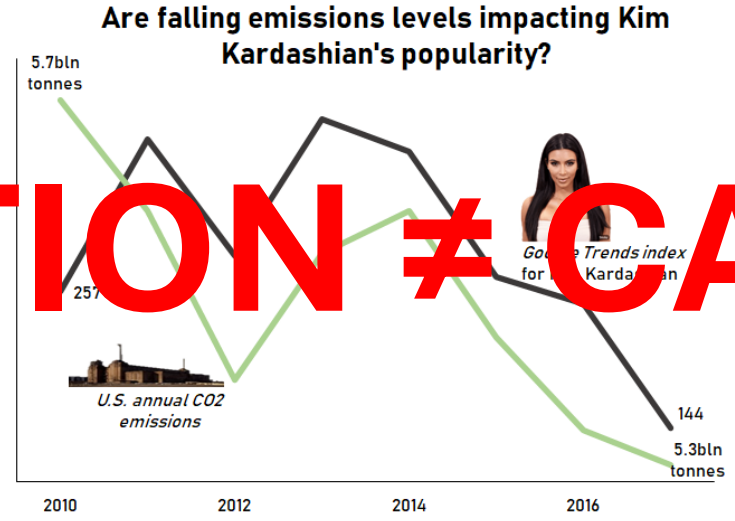


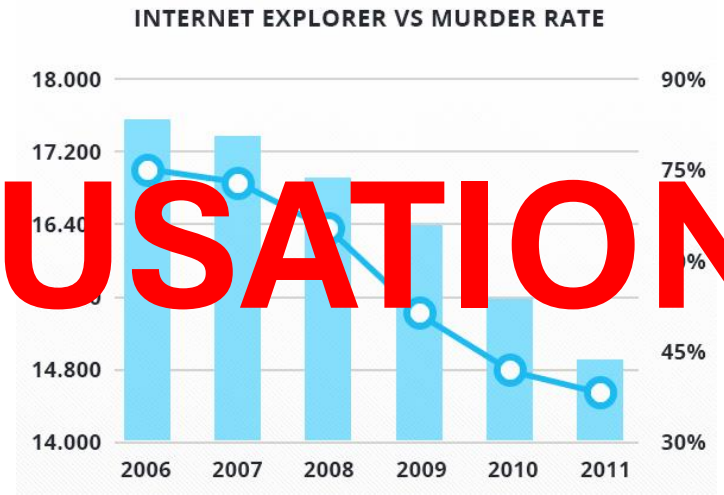
Source: Original work by the author; Mathworks Inc. MRI dataset; BACH: Breast Cancer Histology images.

Interpretability

- Ensuring features are meaningful and interpretable for clinical use.







CORRELATION ≠ CAUSATION

Source: https://medium.com/pythoneers/the-great-debate-correlation-vs-causation-explained-b17e98158760

VIBE PROJECT
Erasmus+
PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS
Silesian University of Technology
U.PORTO
DEX innovation centre

Case study

ECG Feature Extraction for Arrhythmia Detection

Objective: Detect arrhythmias using ECG signal features.

Steps:

1. Preprocessing - Remove noise and artifacts
2. Feature extraction
   - RR Interval (time between consecutive R-wave peaks): Indicates heart rhythm.
   - P-Wave Duration: Duration of atrial depolarization.
   - Frequency Analysis: Apply Fourier transform to detect frequency bands of interest.
3. Feature selection
   - Use filter methods to select features that correlate with arrhythmias.
   - Implement wrapper methods to test feature effectiveness in machine learning models.
4. Classification

VIBE PROJECT
Erasmus+
PÉCSI TUDOMÁNYEGYETEM UNIVERSITY OF PÉCS
Silesian University of Technology
U.PORTO
DEX innovation centre

Conclusion

- Summary: Reviewed feature types, extraction methods, and selection techniques.
- Further Reading:
  - Biomedical Data Science resources
  - Advanced texts on feature engineering in biomedical contexts

Bibliography

1. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer
2. Rangayyan, R. M. (2024). *Biomedical Signal Analysis: Contemporary Methods and Applications*. Wiley-IEEE Press.
3. Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.
4. Dua, S., & Acharya, U. R. (2011). *Data Mining in Biomedical Imaging, Signaling, and Systems*. CRC Press.
5. Saeys, Y., Inza, I., & Larrañaga, P. (2007). *A review of feature selection techniques in bioinformatics*. Bioinformatics, 23(19), 2507–2517.
7. Feature Engineering for Machine Learning – https://towardsdatascience.com/feature-engineering
8. The Great Debate: Correlation vs. Causation – https://medium.com/pythoneers/the-great-debate-correlation-vs-causation-explained-b17e98158760
9. Feature Selection in Python – https://machinelearningmastery.com/tips-for-effective-feature-selection-in-machine-learning/
10. ANOVA Explained – https://www.datacamp.com/tutorial/anova-test